Chemometrics and Intelligent Laboratory Systems 92 (2008) 179–185

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Multilogistic regression by evolutionary neural network as a classification tool to discriminate highly overlapping signals: Qualitative investigation of volatile organic compounds in polluted waters by using headspace-mass spectrometric analysis

César Hervás [a], Manuel Silva [b,*], Pedro Antonio Gutiérrez [a], Antonio Serrano [b]

[a] Department of Computer Science, Albert Einstein Building, University of Cordoba, E-14071 Cordoba, Spain
[b] Department of Analytical Chemistry, Marie-Curie Building (Annex), University of Cordoba, E-14071 Cordoba, Spain

ARTICLE INFO

Article history:
Received 26 November 2007
Received in revised form 12 March 2008
Accepted 13 March 2008
Available online 27 March 2008

Keywords:
Logistic regression
Product unit neural networks
Multi-class pattern recognition
Volatile organic compounds
Headspace-mass spectrometric analysis

ABSTRACT

This work investigates the ability of multilogistic regression models including nonlinear effects of the covariates as a multi-class pattern recognition technique to discriminate highly overlapping analytical signals using a very short number of input covariates. For this purpose, three methodologies recently reported by us were applied based on the combination of linear and nonlinear terms which are transformations of the linear ones by using evolutionary product unit neural networks. To test this approach, drinking water samples contaminated with volatile organic compounds such as benzene, toluene, xylene and their mixtures were classified in seven classes through the very close data provided by their headspace-mass spectrometric analysis. Instead of using the total ion current profile provided by the MS detector as input covariates, the three-parameter Gaussian curve associated to it was used as linear covariates for the standard multilogistic regression model, whereas the product unit basic functions or their combination with the linear covariates were used for the nonlinear models. The hybrid nonlinear model, pruned by a backward stepwise method, provided the best classification results with a correctly classified rate for the training and generalization sets of 100% and 76.2%, respectively. The reduced dimensions of the proposed model: only three terms, namely one initial covariate and two basis product units, enabled to infer interesting interpretations from a chemical point of view.

1. Introduction

Qualitative analysis is increasingly being viewed as an emerging branch of analytical chemistry [1] due to the combination of powerful instrumental techniques, such as chromatography with mass spectrometric or diode-array detector, Fourier-transform infrared spectroscopy, chemical microsensors array, etc. with chemometric methods which expands the possibilities of the identification. "Classification according to specific criteria" is the general definition for qualitative testing [2], which also includes the well-known screening methods used for finding out if a sample contains one or more specific analytes based on a binary response. So, in a broad sense, qualitative analysis is really a simple classification methodology. Several chemometric methods have been used in Analytical Chemistry for qualitative analysis as supervised or unsupervised classification tools. Although a detailed review of these methods is out of the scope of this paper, some choices are factor analysis, cluster analysis (CA), K-Nearest Neighbours (KNN), linear and quadratic discriminant analysis (LDA & QDA), partial least squares discriminant analysis (PLS-DA), and soft independent modelling of class analogy (SIMCA), among others [3,4]. In addition, artificial neural networks (ANNs), with their pattern recognition and modelling capabilities, have currently become powerful classification tools in qualitative chemical analysis [5,6].

Multilogistic regression is a special case of generalized linear model methodology where the assumptions of normality and constant variance of the residuals are not satisfied [7,8]. Multilogistic regression models have demonstrated their accuracy in many classification frameworks often providing classifiers easily interpretable that can be used for the appropriate selection of the adequate model in real supervised learning situations [9–11]. In the last few years, multilogistic regression models have shown their potential as classification and screening tools particularly in clinical analysis, where they have been applied for predicting the probability of suffering a certain disease in terms of the levels of different biomarkers [12–15] or for providing a positive or negative diagnostic over a form of cancer on the basis of different indicators such as serum proteins profile [16,17], concentrations of certain trace elements in bodily fluids [18] and plasma levels of organochlorines [19], among others. In addition, other classification problems have been successfully addressed with multilogistic regression such as the characterization of honey from its mineral content [20], the characterization of pharmaceuticals based on

* Department of Analytical Chemistry, Marie-Curie Building (Annex), Rabanales Campus, University of Cordoba, E-14071 Cordoba, Spain. Tel.: +34 957 212099; fax: +34 957 218614.
E-mail address: qa1sirom@uco.es (M. Silva).

0169-7439/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2008.03.005

the determination of polymorphic purity [21] and the authentication of virgin olive oils of very close geographical origins by near infrared spectroscopy [22].

In this work, the potential of a methodology recently developed by our research group [23] is evaluated as an analytical classification tool for the discrimination between classes that provide highly overlapped signals using a very short number of initial covariates. Specifically, the methodology implies an enhancement of the standard multilogistic regression by including the nonlinear effects of the covariates on the basis of the hybridization of the initial and nonlinear transformed covariates. These nonlinear transformed covariates are constructed with product unit (PU) basis functions given by products of the inputs of a product unit neural network (PUNN) raised up to real powers, which capture the possible strong interactions between the variables. PU functions correspond to a special class of feed-forward neural networks, namely PUNN, introduced by Durbin and Rumelhart [24] and subsequently developed by other authors [25–30]. In this way, standard and product unit basis functions covariates multilogistic regression models were tested [23], including the standard multilogistic regression model (MR) based on the initial covariates and two other MR models using product unit basis functions: the first constructed only on the PU basis functions of the PUNNs (MRPU) and the second with both PUs and initial covariates (MRIPU).

The classification efficiency on the training and the generalization data sets of these models are compared among themselves and also with those provided by classical statistical algorithms such as LDA and QDA. Moreover, recent common classification methods of artificial intelligence, such as support vector machine (SVM) and decision trees, are applied and compared to these MR models. SVM [31,32] is a very popular tool in machine learning that explores the kernel techniques with good geometric explanation, and which usually performs very well in many classification applications; whereas the decision tree algorithm for classification constructs decision trees, where the leaves represent classifications and the branches represent conjunctions of features that lead to those classifications [33,34].

To test this classification tool, a complex analytical pattern recognition problem was investigated, namely the classification of drinking water samples contaminated by the volatile organic compounds (VOCs) benzene, toluene and xylene based on headspace-mass spectrometric (HS-MS) data. The seven classes of water samples contain one of these compounds as well as their binary or ternary mixtures. The complexity of this chemical system is related to highly overlapped MS signals provided by the classes in study. In addition, a data treatment methodology is proposed to extract useful chemical information from the volatile profile provided by the HS-MS based on the decreasing of the number of input parameters, and whose aim is to use as simple multilogistic regression models as possible. So, the number of covariates, used also as inputs to the PUNNs, was estimated by the Levenberg–Marquardt method in the form of a three-parameter Gaussian curve associated with the total ion current profile provided by the MS detector using a similar methodology to the one based on Gaussian and Weibull functions [35–39], previously reported by us. These compounds were also chosen because they belong to the group of most dangerous water pollutants. Their recognition is very important due to differences in toxicity of these compounds and their impact on human health and the environment.

Despite the relatively low number of patterns used for the training and generalization sets, the proposed approach provided good results for the qualitative investigation of these VOCs in polluted waters, and to our knowledge no study on the use of nonlinear transformed covariates multilogistic regression models has to date been reported in this context. In addition, it provides a more quality information than the classical screening methods based on the typical binary response, and it can be extended to others methodologies that insert analytes from a sample directly into a MS, such as membrane introduction MS, among others.

2. Classification method

In multiclassification problems, measurements x_i ($i=1,2,\ldots,k$) are made from a single individual (or object), and individuals are classified into one of J classes on the basis of these measurements. In this paper, the common technique of representing the class levels with a "1-of-J" encoding vector $\mathbf{y}=(y^{(1)},y^{(2)},\ldots y^{(J)})$ is used. Thus, from a training sample defined as $D=\{(\mathbf{x}_n,\mathbf{y}_n);n=1,2,\ldots,N\}$ in which $\mathbf{x}_n=(x_{1n},\ldots,x_{kn})$ is the vector of measurements taking values in $\Omega\subset R^k$ and $\mathbf{y}_n$ is the class level of the nth individual, the query is to find a decision function $C:\Omega\rightarrow\{1,2,\ldots J\}$ based on generalized linear regression models, with new covariates based on PU basis functions for classifying the individuals. A misclassification occurs when a decision rule C assigns an individual (based on measurements vector) to a class j when it is actually coming from a class $l\neq j$.

The logistic regression methods are common statistical tools for modelling discrete response variables; such as multiple, binary, categorical and ordinal responses. In the multiple and categorical cases, the conditional probability that $\mathbf{x}$ belongs to class l verifies:

$$p\left(y^{(l)}=1|\mathbf{x}\right)>0,\ l=1,2,\ldots,J,\ \mathbf{x}\in\Omega \tag{1}$$

and sets the function:

$$f_l(\mathbf{x},\theta_l)=\log\frac{p(y^{(l)}=1|\mathbf{x})}{p(y^{(J)}=1|\mathbf{x})},\ l=1,2,\ldots,J,\ \mathbf{x}\in\Omega \tag{2}$$

where $\boldsymbol{\theta}_l$ is the weight vector corresponding to the class l and $f_J(\mathbf{x},\boldsymbol{\theta}_J)\equiv 0$ considering the J class as the base class. Under a multilogistic regression model, the probability that $\mathbf{x}$ belongs to class l is then given by

$$p\left(y^{(l)}=1|\mathbf{x},\boldsymbol{\theta}\right)=\frac{\exp f_l(\mathbf{x},\boldsymbol{\theta}_l)}{\sum_{j=1}^{J}\exp f_j(\mathbf{x},\boldsymbol{\theta}_j)},\ l=1,2,\ldots,J \tag{3}$$

where $\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_{J-1})$.

An individual should be assigned to the class which has the maximum probability, given the vector measurement $\mathbf{x}$, that is: $C(\mathbf{x})=\hat{l}$, where $\hat{l}=\arg\max_l f_l(\mathbf{x},\hat{\boldsymbol{\theta}}_l)$, for $l=1,\ldots,J$. In this work, a multilogistic regression model developed recently by us based on the combination of linear and nonlinear terms [23] is applied to qualitative analysis. The nonlinear part of the function $f_l(\mathbf{x},\boldsymbol{\theta}_l)$ corresponds to a special class of feed-forward neural networks, namely PUNNs, an alternative to the standard sigmoidal neural networks, which are based on multiplicative nodes instead of additive ones, which have been mainly used for regression problems [24–30].

In a supervised learning process, the components of the weights vector $\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_{J-1})$ are estimated from the training data set D. To achieve the maximum likelihood for the estimation of $\boldsymbol{\theta}$, one can minimize the negative log-likelihood function. The estimation of the vector parameter $\hat{\boldsymbol{\theta}}$ is carried out by means of a hybrid procedure. So, the methodology is based on the combination of an evolutionary algorithm (global explorer) and a local optimization procedure (local exploiter) carried out by the standard maximum likelihood optimization method. The process of estimation of the basis function coefficients is structured in three steps and more specific details about it and the corresponding optimization procedure can be seen in [23].

It is important to remark some characteristics of the first step of the methodology, an evolutionary programming (EP) algorithm that find the weights and the number of PU functions for the MRPU and MRIPU models. The population-based evolutionary algorithm for the architectural design and estimation of real-coefficients was selected on the basis that crossover is not used due to its potential disadvantages in evolving artificial networks [40], although it has common points with other evolutionary algorithms reported in the literature [40–45]. The search begins with an initial population selected randomly. On each

generation of the population is updated using a population-update algorithm. The population is subject to the operations of replication and parametric and structural mutation. More details about the specific characteristics of this evolutionary algorithm are reported elsewhere [28,29,46].

3. Experimental

Experimental data matrix was obtained by using the HS-MS data provided by a set of 63 drinking water samples spiked with individual standards of benzene, toluene or xylene as well as with binary or ternary mixtures of them at concentrations between 5 and 30 μg/l. Table 1 shows the composition of the resulting seven classes. The samples were usually prepared in duplicated as described elsewhere [47] and the total ion current profile provided by the MS detector operated in full scan mode with a range from m/z 50 to 110 at 10.3 scans/s was used as the analytical signal. The experimental design is based on the random distribution of the water samples of each class using 2/3 and 1/3 for the training and generalization sets, respectively (see Table 1).

The Levenberg–Marquardt algorithm was used to estimate the three-parameter Gaussian function associated with the total ion current profile provided by the MS detector. These parameters were defined as follows: $\hat{I}_m$ (maximum abundance), $\hat{t}_m$ (time corresponding to the maximum abundance) and $\hat{B}$(dispersion of the abundance values from I_m). In order to use these parameters as inputs to the assayed PUNNs, they were scaled over the range 0.1 to 0.9. Thus, the new scaled variables were expressed as follows: $\hat{I}_m^*$, $\hat{t}_m^*$ and $\hat{B}^*$. After optimizing the network models, estimations should be de-scaled according to the same procedure.

Table 1
Composition of the training and generalization sample sets

	Training sample set, μg/l			Generalization sample set, μg/l		
	Benzene	Toluene	Xylene	Benzene	Toluene	Xylene
Class 1	5			5		
	10			10		
	15			30		
	15					
	30					
	30					
Class 2		5			5	
		10			30	
		10			30	
		15				
		15				
		30				
Class 3			5			10
			5			15
			10			30
			15			
			30			
			30			
Class 4	5	10		5	5	
	5	30		5	30	
	10	30		30	30	
	15	15				
	15	15				
	30	30				
Class 5	5		10	5		5
	5		30	5		30
	10		30	30		30
	15		15			
	15		15			
	30		30			
Class 6		5	5		10	30
		5	10		30	30
		5	30		30	30
		5	30			
		15	15			
		15	15			
Class 7	5	10	5	5	5	5
	5	10	5	5	30	5
	5	30	5	10	30	10
	15	15	15			
	15	15	15			
	30	30	30			

Multilogistic regression models were fitted to the data obtained by means of a multilogistic regression modelling procedure included in SPSS 12.0 for Windows [48]. The multilogistic regression was performed with a backward conditional method, which selects the most significant covariates. To measure the classifier's performance, the output was compared to the observed outcome of the seven classes, and the correctly classified rate (CCR) for training and generalization sets were obtained, CCR_T and CCR_G, respectively; the CCR being defined as follows:

$$\mathrm{CCR} = \frac{1}{N}\sum_{n=1}^{N} I\left(C(\mathbf{x}_n) = \mathbf{y}_n\right) \tag{4}$$

where $I(\cdot)$ is the zero-one loss function. A good classifier tries to achieve the highest generalization CCR value in a given problem.

The parameters used in the EP algorithm were the same in the two product unit basis functions multilogistic regression models. We consider the same parameter values than those reported in [49] due to their robustness. The exponents w_{ji} and the coefficients β_j^l were initialized in the [−5,5] interval. The size of the population is N=1000 and the maximum number of hidden nodes is m=4. The number of nodes that can be added or removed in a structural mutation is within the [1,2] interval, whereas the number of connections that can be added or removed in a structural mutation is within the [1,6] interval. The number of runs of the EP algorithm was 30 with 200 generations for each run.

Results are compared using LDA, QDA, SVM and the C4.5 tree inducer algorithm for classification. Regarding the classical statistical algorithms, it is well known that if the input variables have a Gaussian distribution and we assume that the variance–covariance matrices are equal, then LDA produces the best Bayes error over the training set, and if the variance–covariance matrices are different the same property is satisfied for QDA. The C4.5 classification tree inducer algorithm is run with the standard options: the confidence threshold for pruning is 0.25, the minimum number of instances per leaf is 2. For pruning, both subtree replacement and subtree rising are considered. The SMO and J48 algorithms are a java implementation of the SVM and C4.5 methodologies, which are part of the Weka machine learning workbench [50] release 3.4.0, using the default parameter values.

4. Results and discussion

Direct sampling MS methods are based on the insertion of the analytes from a sample into a MS using a simple interface with minimal sample preparation and no prior chromatographic separation [51–53]. The recent development of a methodology based on the direct coupling of a headspace sampler with a mass spectrometry detector (HS-MS) has enabled a drastic reduction in analysis time increasing the sample throughput [54]. Generally, data can be obtained by using the mass spectrum that represents the sum of intensities of all the ions detected during the data-acquisition time. Afterwards, it is necessary to extract the information contained in the profile signal and convert it into useful information, which requires the use of chemometric approaches. To date, most applications of HS-MS have focused on qualitative analysis related to quality control in foods [55–58] and environmental pollution [59,60] by applying different classification techniques such as LDA, CA, SIMCA and KNN among others.

As stated above, the goal of this work was to evaluate the potential of a multilogistic regression model recently reported by us [23] in

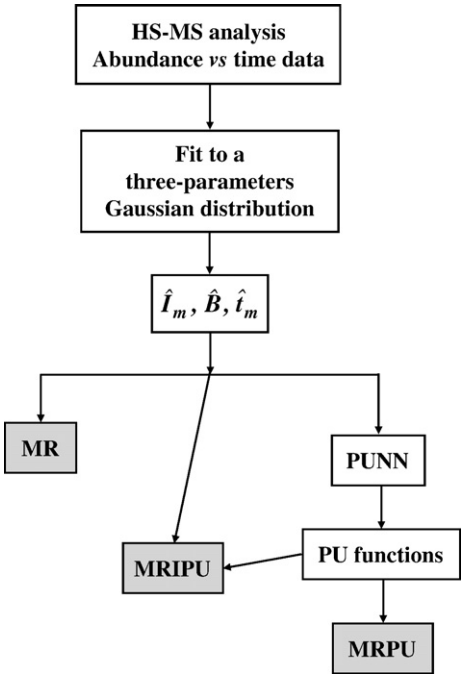

Fig. 1. Flow diagram of the whole chemometric protocol.

qualitative analysis. This methodology, which includes the nonlinear effects of the covariates, has only been applied to benchmark problems. To evaluate the response of the methodology in qualitative analysis, several strategies were merged in this work: 1) the selection of an analytical classification problem in which the classes involved provided analytical signals with a high degree of overlapping, such as the case of the HS-MS profiles obtained in the analysis of drinking waters contaminated by VOCs; 2) the use of a limited number of variables, such as those estimated by the Levenberg–Marquardt method in the form of a three-parameter Gaussian curve associated with the total ion current profile provided by the MS detector; and 3) the use of the PU basis functions as nonlinear terms for the multilogistic regression models. Fig. 1 shows a schematic diagram of the multilogistic regression approaches tested in this paper for the classification of assayed polluted drinking water samples.

4.1. Treatment of the HS-MS data

The volatile profiles provided by the HS-MS instrument corresponding to the seven classes of contaminated drinking waters under study are shown in Fig. 2A, in which a representative profile of each case is included. As can be seen, although the classes containing benzene and toluene give narrower bands (e.g. see peaks 1 and 4) and those containing xylene provide wider ones (e.g. see peaks 3 and 5), the profiles were very similar in general, which makes clear the high overlapping grade of the bands of the different classes. On the other hand, Fig. 2B shows the volatile profile of an un-contaminated and contaminated drinking water (sample 4 in Fig. 2A) in order to evaluate the contribution of the background signal. As can be seen, the signal provided by the analytes can be successfully differentiated from the background resulting in a net signal as shown in Fig. 2C.

The shape of the analyte volatile profile suggests that it should be modelled with a pre-determined function so that its definite parameters can be used as the initial covariates for the MR model, also obtaining the PU basis functions for the MRPU and MRIPU models. This situation provides simpler MR models as well as reduces PUNN complexity and learning time, which is of great practical interest. Taking into account that the resulting analyte volatile profile bands are practically symmetric, the three-parameter Gaussian function was found to be the best choice for modelling HS-MS data. As stated above, the equation corresponding to this function is defined by the following parameters: $\hat{I}_m$ (maximum abundance), $\hat{t}_m$ (time corresponding to the maximum abundance) and $\hat{B}$ (dispersion of the abundance values from $\hat{I}_m$). Fig. 2C shows the fit provided by the three-parameter Gaussian function on the volatile profile obtained in the HS-MS

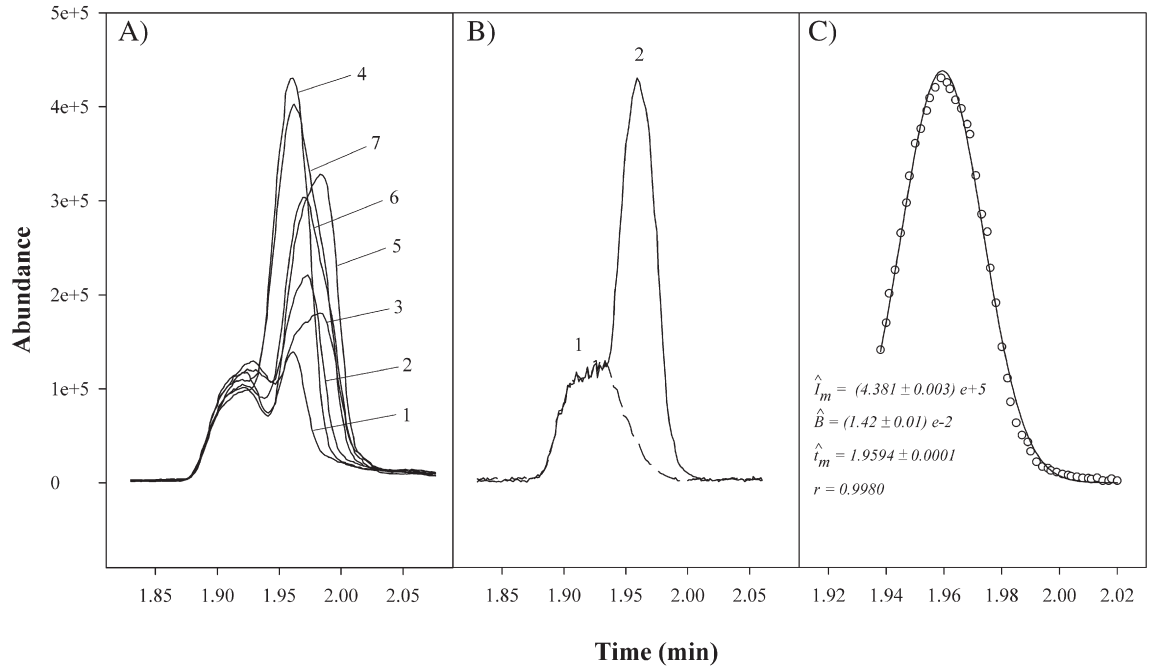

Fig. 2. A) Typical HS-MS profiles corresponding to drinking water samples containing: 1) 15 μg/l of benzene; 2) 15 μg/l of toluene; 3) 10 μg/l of xylene; 4) 10 μg/l of benzene and 30 μg/l of toluene; 5) 5 μg/l of benzene and 30 μg/l of xylene; 6) 15 μg/l of toluene and 15 μg/l of xylene; and 7) 15 μg/l of benzene, 15 μg/l of toluene and 15 μg/l of xylene. B) HS-MS profiles corresponding to: 1) un-contaminated and 2) polluted drinking water with a mixture containing 10 μg/l of benzene and 30 μg/l of toluene. C) HS-MS response fitted to a three-parameter Gaussian distribution. (o) Experimental data and (–) Gaussian curve.

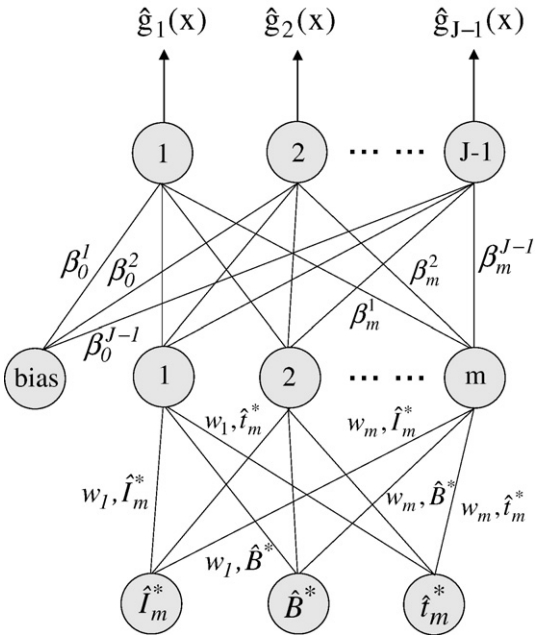

Fig. 3. Functional scheme of the neural network based on PU basis functions. $w_1,\ldots,w_m$ and $\beta_0,\ldots,\beta_m$ are the regression coefficients of the model. Other symbols are defined in the text.

analysis of the drinking water spiked with a binary mixture of benzene and toluene (sample 4 in Fig. 2A). Upon examining both curves and from the estimated statistical parameters, it can be concluded that the three-parameter Gaussian distribution is a fine tool for modelling this kind of HS-MS data.

4.2. Estimation of PU basis functions

Product unit basis functions covariates multilogistic regression models were designed exchanging the classical initial covariates for PU basis functions (MRPU), which are nonlinear themselves, and combining both PU basis functions and the initial covariates (MRIPU). Instead of selecting standard sigmoidal functions, which are based on additive functions, we selected PU basis functions taking into account that PUNNs have an increased information capacity and the ability to form higher-order combinations of inputs [24].

In order to estimate the PU basis functions, the classification ability of different PUNN models, (the general scheme of a PUNN model is shown in Fig. 3) was compared in terms of topology, number of connections, homogeneity (confidence interval) and CCR by using network models with three nodes in the input layer (the three-parameter of the Gaussian function), and six nodes in the output layer (one less than the number of classes as stated in Section 2); thus, 3:5:6 and 3:6:6 architectures were tested.

As shown in Table 2, the classification of the polluted drinking water samples using the best PUNN model, which included five PU basis functions in its hidden layer, is not very satisfactory, and therefore the use of another chemometric approaches such as MR models can be an interesting choice in order to improve the CCR_G values. These PU basis functions also provided useful information from a chemical point of view. In fact, the more relevant PU basis functions (see the discriminant functions, DFs, in Table 2) are $P\hat{U}_2$, $P\hat{U}_4$ and $P\hat{U}_5$ which depend on the initial variables $\hat{t}^*_m$and $\hat{B}^*$, that is, the time corresponding to the maximum abundance of the total ion current profile and the dispersion of the abundance values from it. From these results it is clear that the $\hat{I}^*_m$ values are not necessary for the classification of the contaminated water samples when using PUNNs.

4.3. Evaluation of the multilogistic regression models

As stated above, three multilogistic regression models were tested: MR, MRPU and MRIPU and their features are compared in Tables 3–5; in particular, Table 3 shows the DFs for each model, which depend on the following covariates: $\hat{I}^*_m$, $\hat{B}^*$ and $\hat{t}^*_m$ for the MR; $P\hat{U}_1$ to $P\hat{U}_5$ for MRPU; and $\hat{t}^*_m$, $P\hat{U}_1$ and $P\hat{U}_4$, for MRIPU model, and Table 4 shows the rate of the number of cases that were correctly classified (confusion matrix) for each model. Regarding classification ability (see Table 5), all models provided CCR_T values of 100%; however, the best value for the generalization set (CCR_G) was achieved by using the MRIPU model. Table 5 also shows the results given by other classification algorithms for comparison purposes, such as the classical LDA and QDA algorithms as well as SMO and J48 algorithms implemented for the SVM and C4.5 classification methods. These results justify the use of the MR models proposed in this work to solve the addressed analytical problem. In view of the CCR_T values in Table 5, it can be considered that there is overfitting in MR models; however, if such high percentages of good classification are not achieved in the training process, the CCR_G values are not higher than 66.7%.

By analyzing these results, some chemical interpretations can be drawn from the characteristics of the DFs and the confusion matrix. From the sign and value of the coefficients of the DFs (these functions are related to the discrimination power of one class with respect to the other classes) it can be derived that in general $\hat{B}^*$and $\hat{t}^*_m$ are the most significant covariates for the addressed classification problem. Regarding the proposed MRIPU model, the DFs only depend on the

Table 2
Accuracy, statistical results and PU basis and discriminant functions for the best model obtained by using evolutionary PUNNs (over 30 runs)

PUNN features							
	Connection	CCR_T			CCR_G		
Range topology	Mean ± 1.96 × SD	Mean ± 1.96 × SD	Best	Worst	Mean ± 1.96 × SD	Best	Worst
3:5:6–3:6:6	37.9 ± 8.8	87.9 ± 9.2	97.6	76.2	63.5 ± 12.7	71.4	52.4
PU basis functions							
$P\hat{U}_1$: $(\hat{B}^*)^{-0.40}$			$P\hat{U}_2$: $(\hat{B}^*)^{4.04}(\hat{t}^*_m)^{-1.06}$				$P\hat{U}_3$: $(\hat{I}_m)^{1.91}$
$P\hat{U}_4$: $(\hat{B}^*)^{-1.63}$ $(\hat{t}^*_m)^{2.47}$			$P\hat{U}_5$: $(\hat{B}^*)^{-3.05}$ $(\hat{t}^*_m)^{2.75}$				
Discriminant functions (DF)							
DF_1: 4.64 − 11.4 $P\hat{U}_1$ − 2.44 $P\hat{U}_2$ + 2.29 $P\hat{U}_3$ − 3.61 $P\hat{U}_4$ + 7.30 $P\hat{U}_5$							
DF_2: − 3.29 − 4.04 $P\hat{U}_2$ + 4.09 $P\hat{U}_4$ + 6.12 $P\hat{U}_5$							
DF_3: − 5.89 − 1.80 $P\hat{U}_2$ + 12.9 $P\hat{U}_4$ + 3.20 $P\hat{U}_5$							
DF_4: − 3.38 + 0.52 $P\hat{U}_1$ − 9.25 $P\hat{U}_2$ + 0.97 $P\hat{U}_5$							
DF_5: − 5.28 + 5.29 $P\hat{U}_1$ − 1.28 $P\hat{U}_2$ + 7.69 $P\hat{U}_4$							
DF_6: − 2.18 − 1.57 $P\hat{U}_2$ + 6.15 $P\hat{U}_4$ + 4.09 $P\hat{U}_5$							

Table 3
Discriminant functions (DF) for the MR, MRPU and MRIPU models

MR model

$f_l = \hat{\alpha}_0 + \hat{\alpha}_{l,\hat{t}_m^*}\hat{t}_m^* + \hat{\alpha}_{l,\hat{B}^*}\hat{B}^* + \hat{\alpha}_{l,\hat{t}_m^*}\hat{t}_m^*$ $l = 1,2,\ldots,6$

DF	$\hat{\alpha}_0$	$\hat{\alpha}_{l,\hat{t}_m^*}$	$\hat{\alpha}_{l,\hat{B}^*}$	$\hat{\alpha}_{l,\hat{t}_m^*}$
1	323.3	−65.5	−1074.6	875.9
2	223.1	−37.6	−1188.6	1188.2
3	5.6	210.9	−1269.2	1470.3
4	302.3	49.0	−929.9	661.9
5	57.9	44.9	−258.3	280.3
6	172.0	−41.5	−1041.6	1143.7

MRPU model

$f_l = \hat{\beta}_0 + \hat{\beta}_{l,\hat{PU}_1}\hat{PU}_1 + \hat{\beta}_{l,\hat{PU}_2}\hat{PU}_2 + \hat{\beta}_{l,\hat{PU}_3}\hat{PU}_3 + \hat{\beta}_{l,\hat{PU}_4}\hat{PU}_4 + \hat{\beta}_{l,\hat{PU}_5}\hat{PU}_5$ $l = 1,2,\ldots,6$

DF	$\hat{\beta}_0$	$\hat{\beta}_{l,\hat{PU}_1}$	$\hat{\beta}_{l,\hat{PU}_2}$	$\hat{\beta}_{l,\hat{PU}_3}$	$\hat{\beta}_{l,\hat{PU}_4}$	$\hat{\beta}_{l,\hat{PU}_5}$
1	67.1	−447.3	0.7	−78.9	783.1	−312.6
2	−283.2	−550.0	1.2	−116.9	252.4	543.5
3	129.3	−563.3	−0.4	521.1	1108.2	−625.9
4	109.1	−181.3	−0.1	93.0	415.1	−294.0
5	−78.8	−63.9	−3.9	71.6	−19.2	200.4
6	−896.7	−503.6	2.0	−157.0	−246.9	1695.4

MRIPU model

$f_l = \hat{\alpha}'_{l,\hat{t}_m^*}\hat{t}_m^* + \hat{\beta}'_{l,\hat{PU}_1}\hat{PU}_1 + \hat{\beta}'_{l,\hat{PU}_4}\hat{PU}_4$ $l = 1,2,\ldots,6$

DF	$\hat{\alpha}_{l,\hat{t}_m^*}$	$\hat{\beta}_{l',\hat{PU}_1}$	$\hat{\beta}_{l,\hat{PU}_4}$
1	−394.0	−970.9	1359.2
2	−1238.6	−1235.2	2522.6
3	−2615.3	−1115.7	4164.4
4	666.8	−386.9	−445.4
5	−632.4	−140.2	995.0
6	−2710.6	−864.5	4149.9

covariates $\hat{t}_m^*$ and $\hat{B}^*$ involved in three terms: one linear in $\hat{t}_m^*$ and two nonlinear in $\hat{PU}_1$ and $\hat{PU}_4$, such as $(\hat{B}^*)^{0.40}$ and $(\hat{B}^*)^{-1.63}(\hat{t}^*)^{2.47}$, respectively. Taking into account the sign and value of the respective coefficients of these DFs, it can be inferred that the $\hat{PU}_4$ covariate is the one that exerts a more significant effect in the classification process, except for the class 4 where $\hat{t}_m^*$ is the most relevant. In other words, the interaction between the initial covariates $\hat{B}^*$ and $\hat{t}_m^*$ is the key for the classification of the polluted drinking waters because both are involved in the most relevant term of the DFs and $\hat{B}^*$contributed in a greater extend due to its negative exponent in the PU basis function $\hat{PU}_4 = (\hat{B}^*)^{-1.63}(\hat{t}_m^*)^{2.47}$.

As stated above, Table 4 shows the confusion matrix obtained in the classification of drinking waters by using the three models; it is clear the better performance of the proposed MRIPU model for the classification of these samples. Despite the better results provided by the proposed model, only one sample of the class 2 is correctly classified (the other models also misclassifying these samples). This behaviour can be ascribed to the similar chemical composition of the two unclassified samples in the class 2 (30 μg/l toluene) with other two included in the class 4 (30 μg/l toluene+5 and 30 μg/l benzene, respectively). Taking into account that benzene provides narrower total ion current profile than toluene, its contribution to $\hat{t}_m^*$ and $\hat{B}^*$ in the mixture is less significant than the contribution of toluene, and therefore a slight experimental error in the preparation and/or HS-MS detection of these samples can originate this problem in the correct classification of samples of the class 2.

5. Conclusions

As it has been shown throughout this study, a multilogistic regression model recently reported by us, composed by original covariates and their nonlinear transformations designed by using evolutionary product unit neural networks has demonstrated to be a powerful tool for multi-class pattern recognition in qualitative analysis. Several chemical and chemometric conclusions can be inferred from the results: (i) the improving of the standard MR model by considering the nonlinear effects of the covariates. The ensuing hybrid MR model provided better accurate results for the classification of polluted drinking waters than the other MR alternatives tested and those achieved by the classical discriminant analysis

Table 4
Rate of the number of cases in the generalization set that were classified correctly for the models assayed: (MR, MRPU, MRIPU)

Class predicted/ target	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$	$l=6$	$l=7$
$l=1$	(3, 3, 3)	–	–	–	–	–	–
$l=2$	–	(1, 1, 1)	–	(2, 2, 2)	–	–	–
$l=3$	–	–	(2, 1, 2)	–	–	(1, 2, 1)	–
$l=4$	(1, 0, 0)	–	–	(2, 3, 3)	–	–	–
$l=5$	–	–	–	–	(2, 2, 2)	(1, 1, 1)	–
$l=6$	–	–	(2, 3, 0)	–	–	(1, 0, 3)	–
$l=7$	–	–	–	(0, 0, 1)	(1, 1, 0)	–	(2, 2, 2)

Table 5
Comparison of the quality achieved for the classification of polluted drinking waters using discriminant analysis (LDA and QDA), support vector machine (SVM), model tree algorithm (C4.5), standard multilogistic regression (MR) and multilogistic regression using product unit basis functions methodologies (MRPU and MRIPU)

Algorithm	CCR_T	CCR_G
LDA	90.5	66.7
QDA	100.0	66.7
SVM	86.3	61.9
C4.5	85.7	66.7
MR	100.0	61.9
MRPU	100.0	57.1
MRIPU	100.0	76.2

methodologies such as LDA and QDA and other common classification techniques such as SVM and C4.5. (ii) The good results achieved in spite of the complexity of the analytical problem selected for the evaluation of the models: the classification of seven drinking waters contaminated with benzene, toluene, xylene or their mixtures at μg/l levels using highly overlapping HS-MS data and with a minimum number of patterns in the generalization test. (iii) The simplification of the models by using as initial covariates the three-parameter Gaussian curve associated with the total ion current profile provided by the MS detector, and (iv) the use of an evolutionary algorithm for obtaining the proposed hybrid MR model. Thus, relationships can be inferred from the initial covariates or their PU transformations on the classification process in order to establish the relative influence of each one.

Acknowledgments

The authors gratefully acknowledge the subsidy provided by the Spanish Inter-Ministerial Commission of Science and Technology of the Ministry of Education and Science under the CTQ2007-63962 (Department of Analytical Chemistry, University of Cordoba) and TIN2005-08386-C05-02 (Department of Computer Science, University of Cordoba) Projects. FEDER also provided additional funding. The research of P.A. Gutiérrez has been subsidised by the FPU Predoctoral Program (Spanish Ministry of Education and Science) grant reference AP2006-01746.

References

[1] M. Valcárcel, S. Cárdenas (Eds.), Modern Qualitative Analysis, Trends Anal. Chem., vol. 24, 2005, issue 6.
[2] S.L.R. Ellison, W.A. Hardcastle, Expression of Uncertainty in Qualitative Testing, LGC report LGC/VAM/2002/021, LGC, Teddington, Middlesex, UK, 2002 Available from: www.vam.org.uk.
[3] R.G. Brereton, Chemometrics: Data Analysis for the Laboratory and Chemical Plant, John Wiley & Sons, New York, 2003.
[4] B.K. Lavine, Anal. Chem. 72 (2000) 91R–97R.
[5] G. Dreyfus, Neural Networks: Methodology and Applications, Springer, Heidelberg, 2005.
[6] H.F. Wang, D.Z. Chen, Y.Q. Chen, Chemom. Intell. Lab. Syst. 70 (2004) 23–31.
[7] J.A. Nelder, R.W.M. Wedderburn, J. Roy, Stat. Soc. A 135 (1972) 370–384.
[8] P.Y. McCullagh, J.A. Nelder, Generalized Linear Models, 2nd edn. Chapman & Hall, New York, 1989.
[9] A.M. Aguilera, M. Escabias, M.J. Valderrama, Comput. Stat. Data Anal. 50 (2006) 1905–1924.
[10] S. Dreiseitl, L. Ohno-Machado, J. Biomed. Inform. 35 (2002) 352–359.
[11] D. Hosmer, S. Lemeshow, Applied Logistic Regression, 2nd ed. Willey, New York, 2000.
[12] C.F. Qiu, S. Saito, M. Sakai, K. Ogawa, K. Nagata, M.A. Williams, Clin. Biochem. 39 (2006) 1016–1021.
[13] A. Berdeli, G. Emingil, A. Gurkan, G. Atilla, T. Kose, Clin. Biochem. 39 (2006) 357–362.
[14] Y. Hasui, Y. Hamanaka, N. Okayama, Y. Suehiro, F. Shinozaki, Y. Ueyama, Y. Hinoda, J. Clin. Lab. Anal. 20 (2006) 47–51.
[15] B.L. Mitchell, Y. Yasui, J.W. Lampe, P.R. Gafken, P.D. Lampe, Proteomics 5 (2005) 2238–2246.
[16] M. Rasouli, A. Okhovatian, A. Enderami, Clin. Chem. Lab. Med. 43 (2005) 913–918.
[17] J.N. Li, Z. Zhang, J. Rosenzweig, Y.Y. Wang, D.W. Chan, Clin. Chem. 48 (2002) 1296–1304.
[18] E.A. Hernandez-Caraballo, F. Rivas, A.G. Pérez, L.M. Marco-Parra, Anal. Chim. Acta 533 (2005) 161–168.
[19] A.S. Furberg, T. Sandanger, I. Thune, I.C. Burkow, E.J. Lund, Environ. Monit. 4 (2002) 175–181.
[20] O.M. Hernández, J.M.G. Fraga, A.I. Jiménez, F. Jiménez, J.J. Arias, Food Chem. 93 (2005) 449–458.
[21] M. Blanco, J. Coello, H. Iturriaga, S. Maspoch, C. Pérez-Maseda, Anal. Chim. Acta 407 (2000) 247–254.
[22] E. Bertrán, M. Blanco, J. Coello, H. Iturriaga, S. Maspoch, I. Montoliu, J. Near Infrared Spectrosc. 8 (2000) 45–52.
[23] C. Hervás-Martínez, F.J. Martínez-Estudillo, M. Carbonero-Ruz, Neural Networks, , 2008 accepted, available online 20 January 2008.
[24] R. Durbin, D. Rumelhart, Neural Comput. 1 (1989) 133–142.
[25] D.J. Janson, J.F. Frenzel, IEEE Expert 8 (1993) 26–33.
[26] L.R. Leerink, C.L. Giles, B.G. Horne, M.A. Jabri, Adv. Neural Inf. Process. Syst. 7 (1995) 537–544.
[27] A.P. Engelbrecht, A. Ismail, International Conference on Neural Networks 2002, Proceedings of the Conference, Honolulu, Hawai, 2002.
[28] A.C. Martínez-Estudillo, C. Hervás-Martínez, F.J. Martínez-Estudillo, N. García-Pedrajas, IEEE Trans. Syst. Man Cybern., Part B, Cybern. 36 (2006) 534–545.
[29] A.C. Martínez-Estudillo, F.J. Martínez-Estudillo, C. Hervás-Martínez, N. García-Pedrajas, Neural Netw. 19 (2006) 477–486.
[30] M. Schmitt, Neural Comput. 14 (2001) 241–301.
[31] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, Cambridge, 2000.
[32] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[33] J.R. Quinlan, in: A. Adams, L. Sterling (Eds.), 5th Australian Joint Conference on Artificial Intelligence. Proceedings of the Conference, Singapore, World Scientific, 1995, pp. 343–348.
[34] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.
[35] C. Hervás-Martínez, R. Toledo, M. Silva, J. Chem. Inf. Comput. Sci. 41 (2001) 1083–1092.
[36] C. Hervás-Martínez, M. Silva, J.M. Serrano, E. Orejuela, J. Chem. Inf. Comput. Sci. 44 (2004) 1576–1584.
[37] C. Hervás-Martínez, A.C. Martínez-Estudillo, M. Silva, J.M. Serrano, J. Chem. Inf. Model. 45 (2005) 894–903.
[38] C. Hervás-Martínez, M. Silva, Chemom. Intell. Lab. Syst. 85 (2007) 232–242.
[39] C. Hervás-Martínez, P.A. Gutiérrez, M. Silva, J.M. Serrano, J. Chemom. 21 (2007) 567–577.
[40] P.J. Angeline, G.M. Saunders, J.B. Pollack, IEEE Trans. Neural Netw. 5 (1994) 54–65.
[41] G.F. Miller, P.M. Todd, S.U. Hedge, 3er International Conference on Genetic Algorithms and Their Applications, Proceedings of the Conference, San Mateo, CA, 1989.
[42] X. Yao, Y. Liu, IEEE Trans. Neural Netw. 8 (1997) 694–713.
[43] D.B. Fogel, International Conference on Neural Networks, Proceedings of the Conference, San Francisco, CA, 1993.
[44] N. García-Pedrajas, C. Hervás-Martínez, J. Muñoz-Pérez, Neural Netw. 15 (2002) 1255–1274.
[45] X. Yao, Proceeding of the IEEE 9 (1999) 1423–1447.
[46] C. Hervás-Martínez, F.J. Martínez-Estudillo, Pattern Recogn. 40 (2007) 52–64.
[47] A. Serrano, M. Gallego, M. Silva, Anal. Chem. 79 (2007) 2997–3002.
[48] SPSS, Advanced Models. Copyright 12.0 SPSS Inc., 2003, Chicago, IL.
[49] F.J. Martínez-Estudillo, C. Hervás-Martínez, P.A. Gutiérrez-Peña, A.C. Martínez-Estudillo, S. Ventura, Evolutionary Product-Unit Neural Networks for Classification in Intelligent Data and Automated Learning (IDEAL 2006), Springer, Berlin, 2006, pp. 1320–1328.
[50] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Francisco, USA, 2000.
[51] M.B. Wise, M.R. Guerin, Anal. Chem. 69 (1997) 26A–32A.
[52] S. Bauer, Trends Anal. Chem. 14 (1995) 202–213.
[53] R.T. Short, S.K. Toler, G.P.G. Kibelka, D.T. Rueda Roa, R.J. Bell, R.H. Byrne, Trends Anal. Chem. 25 (2006) 637–646.
[54] J.L. Pérez-Pavón, M. del Nogal-Sánchez, C. García-Pinto, M.E. Fernández-Laespada, B. Moreno-Cordero, A. Guerrero-Peña, Trends Anal. Chem. 25 (2006) 257–266.
[55] F. Peña, S. Cárdenas, M. Gallego, M. Valcárcel, J. Chromatogr. A 1074 (2005) 215–221.
[56] F. Peña, S. Cárdenas, M. Gallego, M. Valcárcel, Anal. Chim. Acta 526 (2004) 77–82.
[57] M.P. Martí, J. Pino, R. Boque, O. Busto, J. Guasch, Anal. Bioanal. Chem. 382 (2005) 440–443.
[58] M.P. Martí, O. Busto, J. Guasch, J. Chromatogr. A 1057 (2004) 211–217.
[59] M. del Nogal-Sánchez, J.L. Pérez-Pavón, C. García-Pinto, M.E. Fernández-Laespada, B. Moreno-Cordero, Anal. Bioanal. Chem. 382 (2005) 372–380.
[60] A. Serrano, M. .Gallego, J. Chromatogr. A 1045 (2004) 181–188.